Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior?

Year: 2020 | Citations: 324 | Authors: Peter Hase, Mohit Bansal

---

Abstract

Algorithmic approaches to interpreting machine learning models have proliferated in recent years. We carry out human subject tests that are the first of their kind to isolate the effect of algorithmic explanations on a key aspect of model interpretability, simulatability, while avoiding important confounding experimental factors. A model is simulatable when a person can predict its behavior on new inputs. Through two kinds of simulation tests involving text and tabular data, we evaluate five explanations methods: (1) LIME, (2) Anchor, (3) Decision Boundary, (4) a Prototype model, and (5) a Composite approach that combines explanations from each method. Clear evidence of method effectiveness is found in very few cases: LIME improves simulatability in tabular classification, and our Prototype method is effective in counterfactual simulation tests. We also collect subjective ratings of explanations, but we do not find that ratings are predictive of how helpful explanations are. Our results provide the first reliable and comprehensive estimates of how explanations influence simulatability across a variety of explanation methods and data domains. We show that (1) we need to be careful about the metrics we use to evaluate explanation methods, and (2) there is significant room for improvement in current methods.